Selecting stimuli parameters for video quality studies based on perceptual similarity distances

Asli Kumcu[a], Ljiljana Platiša[a], Heng Chen[b], Amber Gislason-Lee[c], Andrew G. Davies[c], Peter Schelkens[b], Yves Taeymans[d], and Wilfried Philips[a]

[a]iMinds-IPI, Ghent University, Ghent, Belgium;
[b]iMinds-ETRO, Vrije Universiteit Brussel, Brussels, Belgium;
[c]Division of Biomedical Imaging, University of Leeds, United Kingdom;
[d]Heart Center, Ghent University Hospital, Ghent, Belgium;

ABSTRACT

This work presents a methodology to optimize the selection of multiple parameter levels of an image acquisition, degradation, or post-processing process applied to stimuli intended to be used in a subjective image or video quality assessment (QA) study. It is known that processing parameters (e.g. compression bit-rate) or technical quality measures (e.g. peak signal-to-noise ratio, PSNR) are often non-linearly related to human quality judgment, and the model of either relationship may not be known in advance. Using these approaches to select parameter levels may lead to an inaccurate estimate of the relationship between the parameter and subjective quality judgments – the system's quality model. To overcome this, we propose a method for modeling the relationship between parameter levels and perceived quality distances using a paired comparison parameter selection procedure in which subjects judge the perceived similarity in quality. Our goal is to enable the selection of evenly sampled parameter levels within the considered quality range for use in a subjective QA study. This approach is tested on two applications: (1) selection of compression levels for laparoscopic surgery video QA study, and (2) selection of dose levels for an interventional X-ray QA study. Subjective scores, obtained from the follow-up single stimulus QA experiments conducted with expert subjects who evaluated the selected bit-rates and dose levels, were roughly equidistant in the perceptual quality space - as intended. These results suggest that a similarity judgment task can help select parameter values corresponding to desired subjective quality levels.

Keywords: subjective video quality assessment, multidimensional scaling, image similarity, difference scaling, video compression, laparoscopy, interventional x-ray

1. INTRODUCTION

This work presents a methodology to optimize the selection of multiple parameter levels of an image acquisition, degradation, or post-processing algorithm applied to stimuli intended for evaluation in a subjective image or video quality assessment (QA) study.

Image and video QA studies are typically conducted to optimize and evaluate an imaging system or to evaluate and calibrate objective image quality (IQ) metrics. In either scenario, one or more acquisition or processing parameters will be varied within the study; examples include acquisition settings affecting noise, blur, or contrast levels, or post-processing settings such as compression level or image restoration parameters. The parameter levels of interest (e.g. bit-rate or noise variance) applied to an image are usually not linearly correlated with their perceived quality. For example, a change by one parameter step size (e.g. Δ 1 Mbps) at an intermediate quality level may be more strongly perceived than the same change at extremely high or low quality levels. Thus, parameter levels chosen as a function of the parameter value itself may inadvertently cause increased sampling at the extremes of the quality scale and decreased sampling at intermediate quality levels – levels that are the most likely to be of interest for study. Objective measures, including peak signal-to-noise ratio (PSNR) may also not be suitable for parameter selection as they do not track human perception linearly, must first be calibrated to human data (often modeled as a logistic function[1]), and may be sensitive to content.[2]

Send correspondence to Asli Kumcu, E-mail: asli.kumcu@telin.ugent.be

Incorrectly chosen parameter levels may be the reason some public image and video quality databases suffer from reduced coverage of the quality scale ("Range"), irregularly spaced quality scores ("Uniformity"), and few statistically significant differences between stimuli quality scores ("Discriminability").[3] One possible explanation for the poor performance of databases on these three criteria is an incorrect choice of parameter levels. Some databases choose a non-linearly increasing set of parameter values but do not provide the motivation for the choice of the model[4] or use PSNR to select parameter values.[5,6] Other databases mention a manual tuning procedure carried out by experts or the authors,[7–9] but do not cite a standard QA experimental protocol and analysis method used to conduct the pilot study. Many other QA studies often do not explicitly state how and why certain parameter values were chosen, or use equal step sizes in parameter units. Since a poorly constructed image database may skew the IQ models that rely on these data, it is important to choose parameter values that will generate quality scores with adequate coverage of the quality space, high uniformity, and high discriminability.

The literature is sparse on methodologies for selecting the optimal parameter levels for QA studies. One approach is to conduct a small scale QA study, for example a pilot study using Single or Double Stimulus Continuous Quality Evaluation (SSCQE, DSCQE)[1] methodology. However, absolute judgments, such as those used in image quality scales, may be less reliable than similarity or difference judgments.[10] The use of a small number of subjects, which is often the case in pilot studies, may accentuate the pitfalls of absolute rating scales. Maximum likelihood difference scaling (MLDS),[10,11] which falls in the class of multidimensional scaling (MDS) methods, is a technique which generates a model of the relationship between the parameter of interest and a difference scale from perceived difference judgments conducted on pairs of image pairs (quadruples). While MLDS shows promise as a methodology for selecting parameter levels, it requires a setup for evaluating four stimuli simultaneously. This may present practical difficulties for the evaluation of High Definition (HD) video.

In this paper, we propose a parameter selection (PS) methodology that can be used to select parameter levels such that quality scores are evenly distributed within the desired range. The procedure consists of a paired comparison subjective QA protocol in which a few subjects judge the perceived similarity in quality of two images modified by the parameter of interest. An analysis procedure which employs classical MDS[12] is used to model the relationship between the the perceived similarity distances and the parameter of interest, from which the optimal, perceptually equidistant degradation/processing levels can be chosen. The selected parameter levels from the PS study can then be used in a standard QA study. Experiments conducted with two clinical applications demonstrate this methodology: (1) selection of compression levels for a subjective laparoscopic surgery video experiment, and (2) selection of dose levels for a subjective interventional X-ray experiment. Each application consists of the parameter selection phase, followed by a single stimulus (SS) QA experiment conducted at the selected parameter levels. Section 2 explains the stimuli, protocol, and analysis used in the subjective experiments including the parameter selection methodology in section 2.2.1; results are in section 3; the discussion and limitations of the method are in section 4; concluding remarks are in section 5.

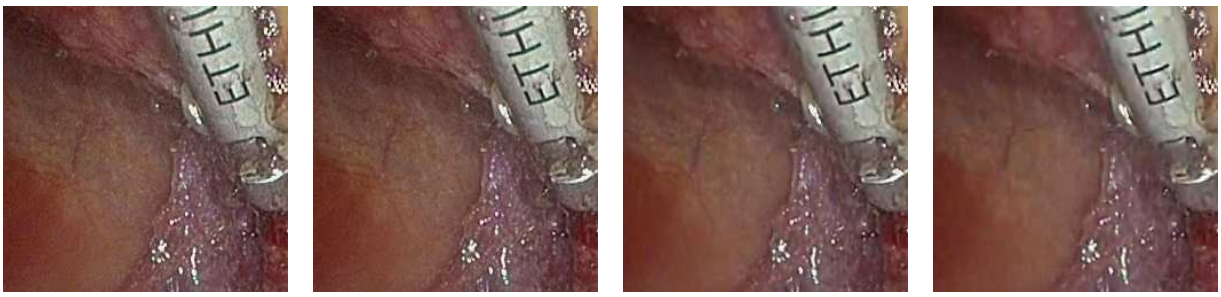
2. METHODS & MATERIALS

Two subjective QA experiments were carried out per application: the paired-comparison PS experiment, followed by the SS experiment conducted at the selected parameter levels. For each application, the stimuli and participant profile are explained, followed by a description of the common protocol and analysis methods.

2.1 Stimuli & Participants

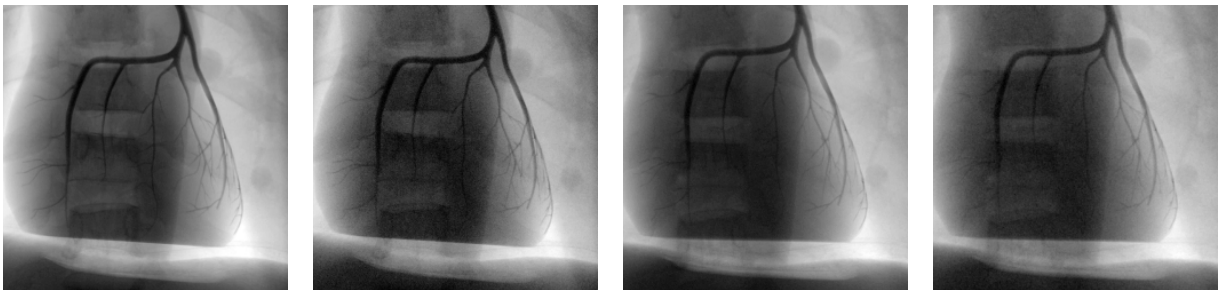
2.1.1 Compression in laparoscopic video

A 10-second scene from a High-Definition (1920 x 1080 pixels) laparoscopic surgery video was extracted and compressed by H.264/AVC at seven bit-rates (20, 8.5, 5, 3.5, 2.5, 1.85, 1.5 Mbps) with parameters optimized for low-latency encoding and decoding, resulting in 8 stimuli (example frames shown in Fig. 1). Details regarding the conversion from the raw image acquisition format to the reference and H.264/AVC compressed sequences may be found in [2]. The PS experiment was conducted on scene D from [2]. We consider two successive compression bit-rates, e.g. 20 and 8.5 Mbps, as being parametrically adjacent. In other applications, parametrically-adjacent values could be, for example, two successive noise or blur levels.



(a) Uncompressed (b) 20 Mbps (c) 2.9 Mbps (d) 1.85 Mbps

Figure 1: Example frames from the reference and three degraded sequences used in the compression experiments (crops of 500x500 are shown)



(a) Standard, 2613 μGy/s (b) Standard, 308 μGy/s (c) Large, 9330 μGy/s (d) Large, 2475 μGy/s

Figure 2: Example frames from the highest dose (reference) and lowest dose sequences acquired with the standard and large cardiac phantoms, used in the dose experiments

A total of twenty-eight pairs of sequences (paired combinations of the reference and seven compressed sequences) were presented to two observers in the parameter selection phase. Both participants were screened for visual defects with a Snellen chart visual acuity test and a digital Farnsworth-Munsell 100 Hue color vision test.

For the SS experiment, four perceptually equidistant compression levels selected from the PS experiment were applied to four laparoscopic surgery scenes and presented to 9 laparoscopic surgeons and 16 non-experts (see [2] for details).

2.1.2 Dose levels in interventional x-ray

A static anthropomorphic chest phantom containing contrast filled coronary arteries (Radiology Support Devices Alderson Phantoms, Long Beach, USA) was imaged on an Allura interventional X-ray system (Philips Healthcare, Best, The Netherlands) with and without 10 cm polymethyl methacrylate (PMMA), to simulate standard and large chest thickness, respectively. Both phantom sizes were imaged at six dose levels (measured as Entrance Skin Dose μGy/s), shown in Table 1. We consider two successive dose levels as being parametrically adjacent. Sequences of 2 seconds were acquired in raw format without any post-processing or proprietary image processing and converted to 8-bit raw RGB avi files (890 x 890 pixels, 15 frames per second) for presentation. Example frames are shown in Fig. 2.

In the parameter selection phase, a total of fifteen pairs of sequences (paired combinations of six dose levels) per phantom type were presented to six interventional cardiologists and cardiac electrophysiologists from Ghent University Hospital. Four doctors scored the standard phantom sequences, and four scored the large phantom sequences. Each subject completed the SS experiment immediately after the PS experiment using the same sequences and parameter levels.

2.2 Protocol & Analysis

The protocol for both the PS and the SS experiments consisted of a brief introduction to the aims of the experiment and explanation of the protocol. Presentation ordering was randomized for all experiments. Three

Table 1: Interventional X-ray image acquisition settings

Acquisition settings	Standard phantom						Large phantom (Standard + 10 cm PMMA)					
kV	70	70	70	70	70	70	85	85	85	85	85	85
mA	100	200	300	400	600	800	300	400	500	600	800	785
ms	5	5	5	5	5	5	7	7	7	7	7	10
Dose (ESD, μGy/s)	308	643	978	1302	1973	2613	2475	3320	4140	4980	6700	9330

ESD: Entrance skin dose (measured). Common settings: 0.1 mm copper filter; source to detector distance (SID): 100 cm for standard phantom, 110 cm for large phantom.

(pairs of) sequences were used for training prior to each experiment.

Sequences were displayed on a 24" surgical display (MDSC–2124, Barco, Kortrijk, Belgium) set at 100% luminance, Gamma display function (2.2 gamma), and 6500K color temperature, with no noise reduction or sharpness post-processing.

2.2.1 Parameter Selection

The two sequences within a paired comparison were presented sequentially on the same monitor using video presentation recommendations from ITU-R.[1] For both clinical applications, subjects evaluated the similarity in quality of each video pair using a continuous scale from 0 (Completely different) to 100% (Exactly the same quality).

The parameter selection analysis procedure consisted of six steps. The procedure is inspired by the "stimulus configuration" step of image quality modeling[13] but does not carry out any transformations of the raw scores.

First, a mean dissimilarity matrix was computed from the raw (dis)similarity scores by taking the mean of the raw similarity scores per pair of stimuli across all subjects. Classical MDS[12] was used to reduce the dimensionality of the dissimilarity matrix, producing a distance matrix. The eigenvalues were used to determine the smallest number of dimensions that accurately reproduced the original distances. The Euclidean distances between parametrically adjacent stimuli were computed from the reduced dimension coordinates, generating a monotonic relationship between *parameter level* versus *perceptual distance*. A best-fit function was fit to the perceived distances, resulting in a model of the relationship between the parameter level and quality difference within the range of parameters tested. This function was used to select parameter levels that were perceptually equidistant. The perceived distance corresponding to each parameter level was compared to the single stimulus quality score, explained in the next section. All analysis was conducted in Matlab (Matlab and Statistics Toolbox Release 2007b, The MathWorks, Inc., Natick, Massachusetts, United States). The *cmdscale* function was used to conduct MDS analysis.

Classical MDS assumes and generates perceptual distances on a ratio scale and requires a complete matrix.[12] This procedures also assumes a monotonic relationship between the processing parameter and perceived quality.

2.2.2 Single Stimulus

Quality was evaluated using the SSCQE[1] methodology. For the laparoscopic video compression experiment, non-experts and surgeons were asked to rate the overall quality ("Quality") of each sequence using a continuous scale from 0 (Poor) to 100% (Excellent quality). The experimental procedure and results are explained in [2].

For the interventional X-ray dose experiment, cardiologists were asked to rate the overall quality of each sequence using a continuous scale from 0 (Poor) to 100% (Excellent quality). They were also asked to rate how well the coronary tree could be visualized using a continuous scale from 0 (Poor) to 100% (Excellent).

A difference opinion score (DOS) – the difference in the quality score between the reference (uncompressed or highest dose acquisition) and the degraded sequence – was computed for each degraded sequence per subject. The mean of the DOS scores - the difference mean opinion score (DMOS) - was fit to the predicted perceived distance

using a separate linear regression for each level of content (4 laparosopic surgery scenes, 2 phantom sizes); values are reported rounded to the nearest integer. All reported ranges and error bars are the 95% confidence interval (CI) of the mean. The coefficient of determination (R^2) of the regressions are reported; values are reported rounded to two significant digits. An R^2 near 1 indicates that the perceived distance function is linearly related to quality preferences and may be used to estimate perceptually equidistant parameter levels. The *differences in DMOS scores* (ΔDMOS) between perceptually adjacent stimuli (e.g. the difference in DMOS scores between 20 and 5.6 Mbps) were also computed and compared to the predicted perceptual difference scores. All analysis was conducted with the R[14] statistical package.

3. RESULTS

3.1 Compression in laparoscopic video

3.1.1 Parameter Selection

The parameter selection analysis for the compression application indicated that the first three eigenvalues of the computed distance matrix explained approximately 97% of the variation. Therefore, the Euclidean distance between parametrically adjacent stimuli was computed using the first three dimensions of the distance matrix. The relationship between dissimilarity distances and compression bit-rates was modeled as a power law function $y = \beta_1 x^{\beta_2}$ with $\beta_1 = 348.8462$ and $\beta_2 = -0.7583864$, shown in Fig. 3a. The R^2 of the fit was 0.99.

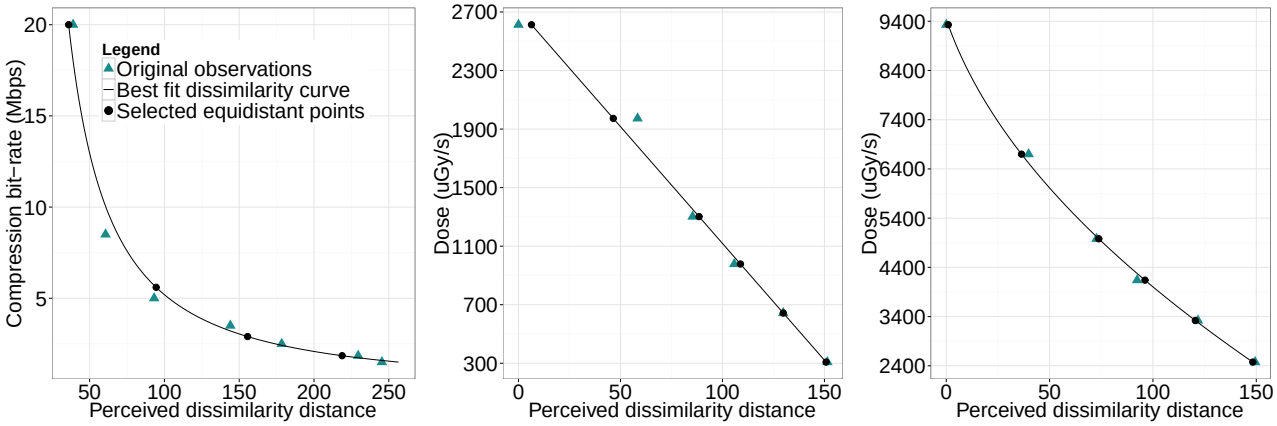
Quality levels between higher bit-rates (less compression) were predicted to be less distinguishable. For example, a 1 Mbps reduction in bit-rate at 20 Mbps translated to a perceived difference of 1.4 units, whereas the difference between 3 and 2 Mbps was predicted to have 54.6 units of perceived difference. That, is the perceived difference between sequences compressed at 3 and 2 Mbps was predicted to be 39 times larger than the perceived difference between sequences compressed at 20 and 19 Mbps.

The goal of the PS experiment was to select four perceptually approximately equidistant compression levels. The highest and lowest quality levels were chosen as anchors as follows: the highest quality level (20 Mbps) was selected such that it was sufficiently similar in quality to the reference sequence, but with some visible compression artifacts. The lowest quality level (1.85 Mbps) was chosen to be sufficiently degraded but not so excessive that the content was visibly destroyed. The two intermediate bit-rates were selected such that the four bit-rates were approximately perceptually equidistant. Intermediate bit-rates that were determined to be exactly equidistant were 5.4 and 2.85 (difference of 61 perceptual units); the intermediate bit-rates that were selected and used in the SS study[2] were slightly different due to the use of two rather than three dimensions to compute the Euclidean distance between parametrically adjacent stimuli. Thus the four bit-rates chosen for the SS study were 20, 5.6, 2.9, and 1.85 Mbps, each bit-rate differing from the adjacent level by 58 to 63 perceptual units; the difference between the uncompressed original and the highest compression bit-rate was 36 perceptual units.

3.1.2 Single Stimulus experiment

The relationship between the perceived distances obtained from the PS experiment - which are roughly equivalent between perceptually adjacent stimuli - and the DMOS scores is shown as a scatter plot in Fig. 4 by subject type and scene.

The DMOS scores for non-experts were approximately evenly spaced, for example a difference of 14–16% $\pm$ 5–8% ΔDMOS between each of the four stimuli in Scene 2, and fit linearly with the perceptual difference predictions (R^2 between 0.97 and 1 for the four scenes). The differences between DMOS scores for surgeons were less equally distributed, ranging approximately 5–26% $\pm$ 8–25% ΔDMOS between the stimuli in the four scenes. A potentially nonlinear relationship between the predicted perceptual distance and the DMOS scores of surgeons may also be observed in Fig. 4, top row (R^2 between 0.91 and 0.97), although it may be noted that the regression falls within the 95% confidence interval (CI). Most DMOS CIs are wider for surgeons ($\pm$ 4–22%, mean 14%) than for non-experts ($\pm$ 5–12%, mean 8%).



(a) Compression bit-rate (b) Dose: Standard Phantom (c) Dose: Large Phantom

Figure 3: Parameter selection analysis results for the three PS experiments: parameter (bit-rate, dose) versus estimated perceived differences

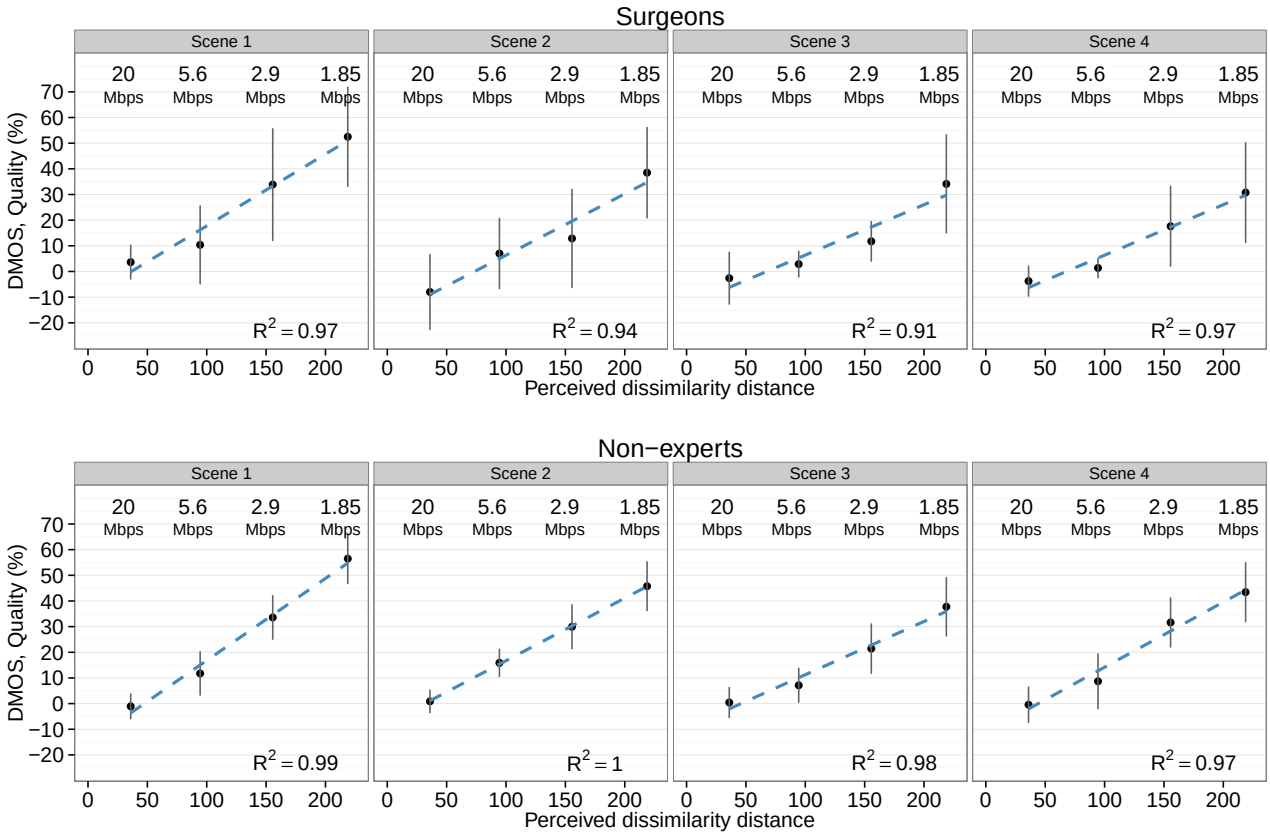


Figure 4: Compressed laparoscopic video results: perceived difference (PS experiment) versus DMOS (SS experiment) by scene (columns) and subject type (surgeons on top row, non-experts bottom row). Error bars are the 95% CI. The dashed blue is the regression line. A DMOS score of 0% indicates no difference between the quality scores of the reference and degraded sequence.

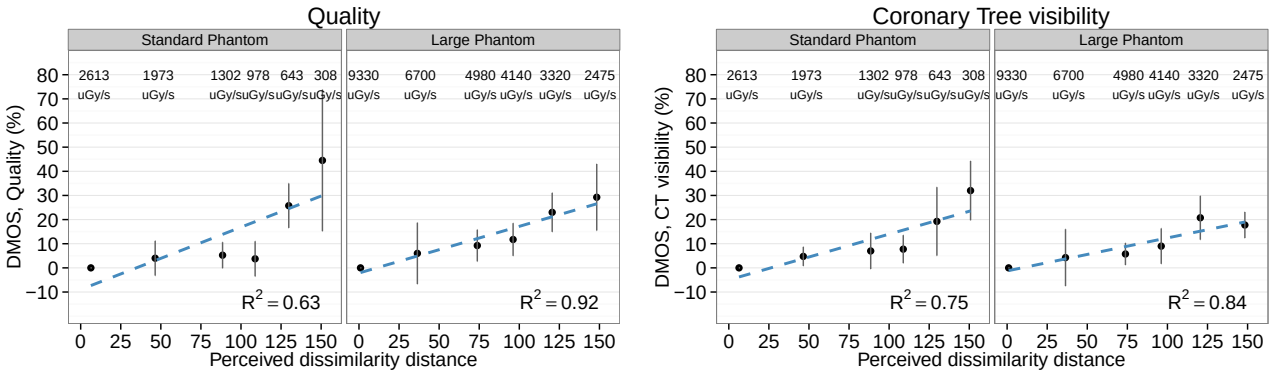


Figure 5: Dose levels in interventional X-ray: perceived difference (PS experiment) versus DMOS (SS experiment). Quality scores on left, visibility of coronary tree on right. Error bars are the 95% CI. The dashed blue is the regression line. A DMOS score of 0% indicates no difference between the quality scores of the reference (highest-dose acquisition) and degraded sequence.

3.2 Dose levels in interventional x-ray

3.2.1 Parameter Selection

Standard phantom The parameter selection analysis for the standard phantom dose application indicated that the first three eigenvalues of the computed distance matrix explained 100% of the variation. Therefore, the Euclidean distance between parametrically adjacent stimuli was computed using the first three dimensions of the distance matrix. The relationship between dissimilarity and dose was modeled as a linear function $y = \beta_1 x + \beta_2$ with $\beta_1 = -0.06266147$ and $\beta_2 = 170.1017$, shown in Fig. 3b. The R^2 of the fit was 0.99.

A range of dose levels were chosen to generate sequences with varying levels of image quality. Since the same dose levels were used in both the PS and SS experiments, we will focus on the relationship between perceived differences and DMOS scores, rather than ΔDMOS.

The difference between dose levels was approximately 40 perceptual units between each adjacent level of the three highest dose levels, and approximately 20 units between each of the lower dose levels.

Large phantom The parameter selection analysis for the large phantom dose application indicated that the first three eigenvalues of the computed distance matrix explained approximately 99% of the variation. Therefore, the Euclidean distance between parametrically adjacent stimuli was computed using the first three dimensions of the distance matrix. The relationship between dissimilarity and dose was modeled as a quadratic function $y = \beta_1 x^2 + \beta_2 x + \beta_3$ with $\beta_1 = 1.9(10^{-6})$, $\beta_2 = -0.04388277$, and $\beta_3 = 245.3101$, shown in Fig. 3c. The R^2 of the fit was 1.

As with the standard phantom study, we will focus on the relationship between perceived differences and DMOS scores, rather than ΔDMOS.

The difference in perceptual units between dose levels were approximately 35 perceptual units between each adjacent level of the three highest dose levels, and 22, 24, and 27 units between each of the lower dose levels.

3.2.2 Single Stimulus experiment

The relationship between perceived difference (from the PS experiment) and DMOS (from the SS experiment) for each level of dose, by phantom size and type of subjective question, is shown as a scatter plot in Fig. 5.

Standard phantom DMOS for the standard phantom were similar at the three highest dose levels below the reference – 1973, 1302, and 978 μGy/s (4–5% ± 5–7% for overall quality and 5–8% ± 4–7% for coronary tree visibility). Differences in DMOS scores (ΔDMOS) between perceptually adjacent stimuli were less than 5%. However there was on average more than 20% loss of quality and visibility at the two lowest dose levels of 643 and 308 μGy/s. Finally, there was very large variation in the quality DOS scores at 308 μGy/s: the minimum and maximum raw quality scores were 25% and 89%, respectively (45% ± 29% DMOS). While the relationship between dose and perceived differences was linear, the relationship between the latter and DMOS appeared to be nonlinear. This behavior can be seen on the first and third plots in Fig. 5 (R^2 0.63 for quality and 0.75 for coronary tree visibility).

Large phantom DMOS for the large phantom varied from 6% ± 13% to 29% ± 14% for quality and 4% ± 12% to 21% ± 9% for coronary tree visibility. The difference in DMOS scores (ΔDMOS) between perceptually adjacent stimuli of the large phantom varied between 3–6% for quality and 2–4% for coronary tree visibility, except for a large loss of quality and visibility at the two lowest dose levels (approximately 11–12% drop in both DMOS values at 3320 μGy/s). The DMOS and perceived differences for the large phantom exhibited a better linear relationship (R^2 of 0.92 for quality and 0.84 for coronary tree visibility) than those of the standard phantom; see the second and fourth plots in Fig. 5.

4. DISCUSSION

The relationship between compression bit-rate and perceived differences was determined to be nonlinear, as might be expected from previous studies,[15] following a power law model. The perceived differences between stimuli at low quality levels (high compression bit-rates) were larger than the differences at high quality levels. The SS experiments conducted with bit-rates chosen in the parameter selection step indicate that the selected stimuli were approximately equally distributed in the Quality (DMOS) space, but fit better to non-experts' than surgeons' scores. As discussed in [2], surgeons scores may have had larger variability compared to non-experts due to their inexperience with subjective quality experiments and the use of quality scales, or differing understanding of quality criteria between subjects. This in turn may have affected the goodness of fit to the perceptual difference predictions. Furthermore, surgeons likely evaluated the quality of surgical video differently than non-experts (including being sensitive to the image content); therefore non-experts may not be suitable predictors of absolute quality scores for surgeons.[2] Nevertheless, these results indicate that the parameter selection methodology may be conducted with a few non-experts to select stimuli parameters that are approximately equally distributed in the quality space, which may then be judged in a QA study with surgeons or other clinicians as expert subjects.

In the dose experiments, a linear relationship between dose and perceived differences was found for the standard phantom. However, it was not able to well predict the distribution of the DMOS scores. The lack of fit may have been affected by the range of the scoring scale used by the cardiologists in the SS experiment: some subjects rated the lowest dose sequences very low in quality, whereas others scored them as moderately low (the 95% CI range was 58% DMOS). For the large phantom, the relationship between dose and perceived differences was found to be nonlinear and best modeled by a quadratic function with a very small coefficient for the quadratic term. The correlation between perceived differences and DMOS scores was moderately linear, lending promise to the use of the methodology used in this study to predict perceptually equidistant dose levels. The values and confidence intervals of the DMOS scores were similar for the quality and the coronary tree visibility tasks, suggesting a correlation between the interpretation or perception of the two tasks.

Normalization of the scores may have compensated for the large variances in scores. However, this would remove any variation in their scores due to differences stemming from factors such as expertise, medical training, or adaptation to images generated from a particular manufacturer's scanner brand, version, or dose/quality optimization setting. Additional studies are needed to reveal whether the differences in quality perceptions of clinicians are simply due to differences in the use of the scale, or differences due to other factors. There were too few subjects in the present study to correlate their use of the scale with their expertise, background, or expectations. Changes to the protocol, such as increasing the number of training sequences, may have the potential to improve the precision of the scores. In the future, we propose to use Thurstone modeling[16] to

transform subjects' scores prior to computing the perceived differences function, rather than using the mean of the dissimilarity scores, in order to account for variability in the use of the similarity scale.

These experiments were restricted to the study of a single parameter that was monotonically related to quality perception. In both applications, only one parameter was varied (bit-rate, tube current) while all other imaging and acquisition parameters were held constant. In device development and clinical use, more than one parameter or processing algorithm may be varied, resulting in a potentially complex relationship with perceived quality. In addition, some imaging parameters may not be monotonically related to quality. For example, while contrast threshold monotonically increases with tube current and decreases with patient thickness,[17] post-processing algorithms such as spatial or temporal filtering may not exhibit a monotonic relationship with perceived quality. Further study is needed to determine how best to model complex relationships in a reduce-dimensionality perceptual space; a two-step MDS approach,[13] which relates dissimilarity scores to specific image quality attributes, may be used to evaluate multidimensional parameter spaces.

Finally, the conclusions of this study are limited to subjective quality preferences, not task performance. As objective measurement of image quality is the gold standard in medical applications, a task-based evaluation is recommended for both applications. It would be of interest to study how, for example, the time to conduct a procedure, or the ability to detect a lesion, relates to the parameter space in medical video applications.

5. CONCLUSION

In this study we present a formal framework for selecting optimal stimuli parameter levels. Pairs of stimuli are judged in terms of similarity and analyzed with classical MDS to build a model between perceived differences and the studied parameter levels. This methodology ensures that the perceived quality scores of the selected parameter levels are evenly distributed within the quality range of interest, to avoid the selection of parameter levels which cluster at the extremes of the quality scale. Experiments conducted for two medical applications – compression bit-rate selection in laparoscopic video and dose selection in interventional X-ray – demonstrate that a paired-comparison, quality similarity judgment task can assist the selection of optimal parameter levels for a subjective QA study. Subjective scores from follow-up single stimulus QA experiments, conducted with the selected bit-rates and dose levels, were roughly equidistant in the perceptual quality space - as intended. In the laparoscopic surgery application, perceptual distances obtained from two non-expert subjects highly predicted the distribution of quality scores from sixteen non-experts and moderately predicted the distribution of nine surgeons' scores. Results from the X-ray dose application indicated moderate predictive performance between perceived similarity and quality for the same four observers; differences in the use of the scale between subjects or insufficient training in the experiment may have resulted in inconsistent scores.

Future work will consider more robust statistical approaches to account for inter-subject variability in the use of the similarity scale. In addition, modeling parameters that have a non-monotonic relationship with perceived quality will be addressed. Finally, a subjective QA study is not a replacement for assessment of task performance, the gold standard for QA in the medical imaging domain. A task-based QA study may show a different relationship between parameter, perceived differences, and image quality. Further research is required to determine the relationship between perceived differences and task-based image QA scores.

6. ACKNOWLEDGMENTS

We would like to thank the six doctors who participated in the interventional X-ray experiment.

Parts of this work were performed within the Telesurgery project (co-funded by iMinds, a digital research institute founded by the Flemish Government; project partners are Unilabs Teleradiology, SDNsquare and Barco, with project support from IWT) and the PANORAMA project (co-funded by grants from Belgium, Italy, France, the Netherlands, the United Kingdom, and the ENIAC Joint Undertaking).

REFERENCES

[1] BT.500-13, I.-R. R., "Methodology for the subjective assessment of the quality of television pictures." International Telecommunications Union (January 2012).

[2] Kumcu, A., Bombeke, K., Chen, H., Jovanov, L., Platiša, L., Luong, H., Van Looy, J., Van Nieuwenhove, Y., Schelkens, P., and Philips, W., "Visual quality assessment of H.264/AVC compressed laparoscopic video," in [*SPIE Medical Imaging, Proceedings*], *Proc. SPIE* **9037**, 90370A–90370A–12 (2014).

[3] Winkler, S., "Analysis of public image and video databases for quality assessment," *Journal on Selected Topics in Signal Processing* **6**, 616–625 (Oct. 2012).

[4] Zhang, F., Li, S., Ma, L., Wong, Y. C., and Ngan, K. N., "IVP subjective quality video database," tech. rep., http://ivp.ee.cuhk.edu.hk/research/database/subjective/ (2011).

[5] De Simone, F., Naccari, M., Tagliasacchi, M., Dufaux, F., Tubaro, S., and Ebrahimi, T., "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," in [*International Workshop on Quality of Multimedia Experience, QoMEX*], 204–209 (July 2009).

[6] Ponomarenko, N., Ieremeiev, O., Lukin, V., Jin, L., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., and Kuo, C.-C. J., "A new color image database TID2013: Innovations and results," in [*Advanced Concepts for Intelligent Vision Systems*], Blanc-Talon, J., Kasinski, A., Philips, W., Popescu, D., and Scheunders, P., eds., *Lecture Notes in Computer Science* **8192**, 402–413, Springer International Publishing (2013).

[7] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K., "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing* **19**(6), 1427–1441 (2010).

[8] Sheikh, H., Sabir, M., and Bovik, A., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing* **15**, 3440–3451 (Nov 2006).

[9] Péchard, S., Pépion, R., and Le Callet, P., "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in [*International Workshop on Image Media Quality and its Applications, IMQA*], 6 (Sept. 2008).

[10] Charrier, C., Maloney, L. T., Cherifi, H., and Knoblauch, K., "Maximum likelihood difference scaling of image quality in compression-degraded images," *J. Opt. Soc. Am. A* **24**, 3418–3426 (Nov 2007).

[11] Maloney, L. T. and Yang, J. N., "Maximum likelihood difference scaling," *Journal of Vision* **3**(8) (2003).

[12] Young, F. W., [*Encyclopedia of Statistical Sciences*], vol. 5, ch. Multidimensional scaling, 649–659, New York: Wiley (1985).

[13] Martens, J., "Multidimensional modeling of image quality," *Proceedings of the IEEE* **90**, 133–153 (Jan 2002).

[14] Team, R. C., *R: A Language and Environment for Statistical Computing, v3.0.2*. R Foundation for Statistical Computing, Vienna, Austria (2013).

[15] Ou, Y.-F., Liu, T., Zhao, Z., Ma, Z., and Wang, Y., "Modeling the impact of frame rate on perceptual quality of video," in [*15th IEEE International Conference on Image Processing (ICIP)*], 689–692 (Oct 2008).

[16] Martens, J.-B., "Interactive statistics with Illmo," *ACM Trans. Interact. Intell. Syst.* **4**, 4:1–4:28 (Apr. 2014).

[17] Dragusin, O., Smans, K., Jacobs, J., Inal, T., and Bosmans, H., "Evaluation of the contrast-detail response of a cardiovascular angiography system and the influence of equipment variables on image quality," *Proc. SPIE* **6913**, 69134R–69134R–12 (2008).